BLAST

From Wikipedia, the free encyclopedia

In bioinformatics, **B**asic **L**ocal **A**lignment **S**earch **T**ool, or **BLAST**, is an algorithm for comparing biological sequences, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences. A *BLAST search* enables a researcher to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold. For example, following the discovery of a previously unknown gene in the mouse, a scientist will typically perform a BLAST search of the human genome to see if human beings carry a similar gene; BLAST will identify sequences in the human genome that resemble the mouse gene based on similarity of sequence.

BLAST

Developer:	Altschul S.F., Gish W., Miller E.W., Lipman D.J., NCBI
Latest release:	2.2.15 /
OS:	UNIX, Linux, Mac, MS-Windows
Use:	Bioinformatics tool
Licence:	Public Domain
Website:	[1] (ftp://ftp.ncbi.nlm.nih.gov/blast/)

Contents

Background

BLAST is one of the most widely used bioinformatics programs, probably because it addresses a fundamental problem and the algorithm emphasizes speed over sensitivity. This emphasis on speed is vital to making the algorithm practical on the huge genome databases currently available, although subsequent algorithms can be even faster.

Examples of other questions that researchers use BLAST to answer are

- Which bacterial species have a protein that is related in lineage to a certain protein whose amino-acid sequence I know?
- Where does the DNA that I've just sequenced come from?
- What other genes encode proteins that exhibit structures or motifs such as the one I've just determined?

BLAST is also often used as part of other algorithms that require approximate sequence matching.

The BLAST algorithm and the computer program that implements it were developed by Stephen Altschul, Warren Gish, David Lipman at the U.S. National Center for Biotechnology Information (NCBI), Webb Miller at The Pennsylvania State University, and Gene Myers at the University of Arizona . It is available on the web at [2] (http://www.ncbi.nlm.nih.gov/BLAST/). Alternative implementations are available at [3] (http://blast.wustl.edu/) (WU-BLAST) and [4] (http://www.fsa-blast.org/) (FSA-BLAST).

The original paper "Altschul, SF, W Gish, W Miller, EW Myers, and DJ Lipman. Basic local alignment search tool. J Mol Biol 215(3):403-10, 1990." was the most highly cited paper published in the 1990s.

Input/Output

Input and Output, complies to the FASTA format

Algorithm

To run, BLAST requires two sequences as input: a query sequence (also called the target sequence) and a sequence database. BLAST will find subsequences in the query that are similar to subsequences in the database. In typical usage, the query sequence is much smaller than the database, e.g., the query may be one thousand nucleotides while the database is several billion nucleotides.

BLAST searches for high scoring sequence alignments between the query sequence and sequences in the database using a heuristic approach that approximates the Smith-Waterman algorithm. The exhaustive Smith-Waterman approach is too slow for searching large genomic databases such as GenBank. Therefore, the BLAST algorithm uses a heuristic approach that is slightly less accurate than Smith-Waterman but over 50 times faster. The speed and relatively good accuracy of BLAST are the key technical innovation of the BLAST programs and arguably why the tool is the most popular bioinformatics search tool.

The BLAST algorithm can be conceptually divided into three stages.

- In the first stage, BLAST searches for exact matches of a small fixed length W between the query and sequences in the database. For example, given the sequences AGTTAC and ACTTAG and a word length W = 3, BLAST would identify the matching substring TTA that is common to both sequences. By default, W = 11 for nucleic seeds.
- In the second stage, BLAST tries to extend the match in both directions, starting at the seed. The ungapped alignment process extends the initial seed match of length W in each direction in an attempt to boost the alignment score. Insertions and deletions are not considered during this stage. For our example, the ungapped alignment between the sequences AGTTAC and ACTTAG centered around the common word TTA would be:

```
..AGTTAC..
  | |||
..ACTTAG..
```

If a high-scoring ungapped alignment is found, the database sequence is passed on to the third stage.

- In the third stage, BLAST performs a gapped alignment between the query sequence and the database sequence using a variation of the Smith-Waterman algorithm. Statistically significant alignments are then displayed to the user.

An extremely fast but considerably less sensitive alternative to BLAST that compares nucleotide sequences to the genome is BLAT (**B**last **L**ike **A**lignment **T**ool). A version designed for comparing multiple large genomes or chromosomes is BLASTZ. Also there is another well-known software called PatternHunter (http://www.bioinformaticssolutions.com/products/ph/index.php) which produces significantly better sensitivity results than BLAST at the same speed or very similar sensitivity results at a much faster speed.

Parallel BLAST

Parallel BLAST versions are implemented using MPI, Pthreads and are ported on various platforms including Windows,Linux, Solaris, OSX, and AIX. Popular approaches to parallelize BLAST include query distribution, hash table segmentation, computation parallelization, and database segmentation(partition).

Program

The BLAST program can either be downloaded and run as a command-line utility "blastall" or accessed for free over the web. The BLAST web server, hosted by the NCBI, allows anyone with a web browser to perform similarity searches against constantly updated databases of proteins and DNA that include most of the newly sequenced organisms.

BLAST is actually a family of programs (all included in the blastall executable). The following are some of the programs, ranked mostly in order of importance:

- **Nucleotide-nucleotide BLAST (blastn)**: This program, given a DNA query, returns the most similar DNA sequences from the DNA database that the user specifies.
- **Protein-protein BLAST (blastp)**: This program, given a protein query, returns the most similar protein sequences from the protein database that the user specifies.
- **Position-Specific Iterative BLAST (PSI-BLAST)**: One of the more recent BLAST programs, this program is used for finding distant relatives of a protein. First, a list of all closely related proteins is created. Then these proteins are combined into a "profile" that is a sort of average sequence. A query against the protein database is then run using this profile, and a larger group of proteins found. This larger group is used to construct another profile, and the process is repeated.
  By including related proteins in the search, PSI-BLAST is much more sensitive in picking up distant evolutionary relationships than the standard protein-protein BLAST.
- **Nucleotide 6-frame translation-protein (blastx)**: This program compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.
- **Nucleotide 6-frame translation-nucleotide 6-frame translation (tblastx)**: This program is the slowest of the BLAST family. It translates the query nucleotide sequence in all six possible frames and compares it against the six-frame translations of a nucleotide sequence database. The purpose of tblastx is to find very distant relationships between nucleotide sequences.
- **Protein-nucleotide 6-frame translation (tblastn)**: This program compares a protein query against the six-frame translations of a nucleotide sequence database.
- **Large numbers of query sequences (megablast)**: When comparing large numbers of input sequences via the command-line BLAST, "megablast" is much faster than running BLAST multiple times. It basically concatenates many input sequences together to form a large sequence before searching the BLAST database, then post-analyze the search results to glean individual alignments and statistical values.

See also

- Needleman-Wunsch algorithm
- Sequence alignment
- Sequerome
- Smith-Waterman algorithm

External links

- NCBI-BLAST website (http://www.ncbi.nlm.nih.gov/BLAST/)

- NCBI-BLAST Tutorial (http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/tut1.html)
- WU-BLAST (http://blast.wustl.edu/) - The original gapping BLAST with statistics, developed and maintained by Warren Gish at Washington University in St. Louis
- FSA-BLAST (http://www.fsa-blast.org/) - A new, faster but still accurate version of NCBI BLAST based on recently published algorithmic improvements
- NBIC mpiBLAST (http://services.nbic.nl:4080/bb/cgi-bin/bb_login.cgi) - Netherlands Bioinformatics Centre, running mpiBLAST
- PatternHunter (http://www.bioinformaticssolutions.com/products/ph/index.php) - An alternative software which provides similar functionality to BLAST while claiming increased speed and sensitivity
- Parallel BLAST (http://www-users.cs.umn.edu/~rangwala/final_bglBLAST.pdf) - A dual scheduling BLAST tested on the Blue Gene/L
- BLAST HOWTO (http://wikiomics.org/wiki/BLAST) at the Wikiomics bioinformatics wiki (http://wikiomics.org/)
- A/G BLAST (http://developer.apple.com/darwin/projects/blast/) - Implementation for PowerPC G4/G5 processors and Mac OS X, from Apple Computer's Advanced Computation Group and Genentech.
- STRAP (http://3d-alignment.eu/) The protein workbench STRAP (http://www.charite.de/bioinf/strap/) contains a comfortable BLAST front-end with a cache for BLAST results

Categories: Articles with unsourced statements | Bioinformatics | Computational phylogenetics | Laboratory software